JOURNAL OF THE EXPERIMENTAL ANALYSIS OF BEHAVIOR 1979, 31, 433-440 NUMBER 3 (MAY)

HOW TO LOOK AT DATA: A REVIEW OF JOHN W. TUKEY'S EXPLORATORY DATA ANALYSIS[1]

RUSSELL M. CHURCH

BROWN UNIVERSITY

Here is a book about data analysis that should be fascinating to many readers of the *Journal of the Experimental Analysis of Behavior.* It is quite different from any standard textbook of statistics: It does not deal with testing hypotheses and establishing confidence intervals. Instead, it provides techniques and advice about how to explore data. The approach is quite compatible with the practices of many readers of *JEAB.*

The book describes many specific methods and a general approach. The approach of exploratory data analysis is described as being detective in character. It is a search for clues. Some of the clues may be misleading, but some will lead to discoveries. After the appearances are described, other techniques can be employed for purposes of confirmation, but this book deals only with the exploratory phase. It is necessary to discover facts before they can be confirmed.

The author, a distinguished statistician, clearly expresses his values. He favors simplicity because simple statements are clear. He particularly features clear visual displays of quantitative facts. He likes precision because a more precise statement contains more information than a less exact statement. For example, it is far better to be able to say some response measure is a linear function of a particular stimulus variable than to say it increases with the stimulus variable. He likes flexibility of approach because it is seldom clear exactly which methods will best achieve the goals of the data analyst, and sometimes different analyses of the same data reveal different aspects of it. He favors depth of analysis. It is always good to look at the residuals; for example, after a linear function is found between a stimulus measure and a response measure, to see if there are systematic deviations from the linear function.

Another of Tukey's values is accuracy, but within reasonable limits. A misplaced decimal point may be serious, but a misplaced digit may not. He proposes methods for checking the accuracy of calculations. In most published reports it is not possible to check the accuracy of data analysis. This is unfortunate since one may suspect that the accuracy is often far from perfect. Finally, he values replicability of summary observations in situations containing occasional aberrant observations.

Most investigators would agree with these sentiments. But an analyst can be flexible, and produce clear, precise, deep, accurate, replicable results only if he has a sufficient number of methods at his disposal. This book describes the methods. It deals with such topics as frequency distributions, measures of central tendency and variability, scale transformations, graphical displays of a single variable and of relationships, smoothing techniques, and analysis of tables. The author obviously enjoyed writing the book, and the treatment of each of the topics is original.

Graphs

Tukey's approach to data analysis is highly visual, and he has numerous suggestions for graphical displays. Graphs are used for many different purposes. They can be used to store quantitative data, to communicate conclusions, or to discover new information. Some types of plots are better for one purpose, and some are better for another. For example, if one uses a graph to store numbers, it is useful to have many markings on the axes, but if one is interested in seeing the form of a relationship, many numbers on the axes distract attention. For looking at the data, Tukey favors

[1]Reading, Mass.: Addison-Wesley, 1977. Pp. 688, $16.95. Reprints of this review may be obtained from Russell M. Church, Walter S. Hunter Laboratory of Psychology, Brown University, Providence, Rhode Island 02912.

paper without distracting grid marks, few numbers on the axes, and fitted lines rather than adjacent points connected with straight-line segments. Tukey particularly emphasizes the value of graphs for discovery. Most of the displays may never be seen in a published report, but they are not designed for publication. Many psychological investigators probably graph only the data they plan to include in an article. This is a mistake. Graphical techniques can be important in the discovery phase of an analysis.

My examination of volumes of *JEAB* at 5-year intervals (1958, 1963, 1968, 1973, and 1978) revealed few obvious trends in type of graphs. There was a reduction in the percentage of articles with one or more cumulative records (from 68% in 1958 to 9% in 1978), and there was an increase in the number of articles with straight lines fitted to the data (from 5% in 1958 to 21% in 1978). This increase in straight-line fits probably reflected the increase in matching-law and signal-detection theory studies, both of which conventionally use straight lines. I had assumed that the level of analysis would be far more advanced in 1978 than in 1958, but some articles in the first volume of *JEAB* used methods of data presentation that would be considered excellent by current standards. These include, but are not necessarily limited to, the articles by Blough (1958), Clark (1958), and Mechner (1958). Some figures in *JEAB* describe apparatus or procedure; most describe results. The figures that describe results can usually be classified into one of the following categories:

1. Representations of frequency distributions of a response measure (histograms and polygons).

2. A measure of response as a function of time, trials, or sessions. Examples are obvious, and they include the output of cumulative recorders and polygraphs, as well as response rate as a function of session, or of time within session.

3. A scatter plot of two response measures. The standard matching-law plot of proportion of responses as a function of proportion of reinforcements received is one example. The ROC curve of signal detection theory is another.

4. A measure of response as a function of an experimenter-controlled variable. The standard generalization gradient is an example, with response rate plotted as a function of wavelength. A psychophysical function is another example.

This book describes new methods of dealing with all of these types of figures.

Are Tukey's graphical methods really better than a cumulative record? The cumulative record is one way to report all of the data from an experiment directly, without distortion. One difficulty is that the record may be so long that it is necessary to select some "representative" subset of the data to present, and the rules for choosing the typical example are not usually well defined. Is it a random one, the best, a prototype, an average, a haphazard choice, etc.? And which session is chosen for display (random, best, prototype, average, haphazard, etc.)? The duration of the displayed cumulative record divided by the total experiment time is usually trivially small. Even a complete cumulative record does not really show everything, because the chart was moved at some fixed speed. A rather different picture might well emerge if the chart speed were increased or decreased substantially. Finally, since no general quantitative ways have been developed to categorize the records, data cannot be combined. This means the reader must do his own data analysis, usually (as noted above) with an inadequate amount of data.

The Distribution of a Single Variable

Tukey's novel "stem and leaf" method of constructing frequency distributions is an improvement over the standard tally method: It is easier to check, and it is easier to use to find measures of central tendency and variability. The measures he proposes involve no arithmetic, only counting. He regularly deals with the median as a measure of central tendency and with the interquartile range as a measure of variability. These have the important feature of being relatively unaffected by occasional observations that deviate substantially from most of the observations and that cannot be explained. In contrast, the mean and the standard deviation can be markedly affected by these occasional outliers.

In *JEAB,* the typical representation of a frequency distribution is a histogram. When looking at a frequency distribution, one should note its height, where it is centered, how spread out it is, whether it is asymmetric, and whether there are any discontinuities. Very

often it is more convenient to look at some transform of the original variable. If the distribution is far from symmetrical (as, for example, reaction time or response rate), one end of the distribution will be too crowded to permit careful inspection. A logarithmic or even a reciprocal transformation could make the data easier to examine.

Transformations may also be employed to make the distribution more symmetrical or to equate the variability of different distributions. For example, the mean and standard deviations of a measure are often linearly related on the original scale but not on the logarithmic scale. It is far more informative to say that the variability is the same on a log scale than to say it is different on an arithmetic scale.

When comparing several frequency distributions, I particularly like what Tukey calls the "box-and-whisker" plots that show medians, quartiles, and two more extreme values in a format that is easy to grasp quickly. In *JEAB* it is uncommon to display visually a measure of variability and, I believe, unprecedented to present two such measures. It is, however, a policy to provide some measure of variability, as well as a measure of central tendency, whenever data are combined (Zeiler, 1977).

Fitting a Straight Line

When there is a linear relationship between two variables, many psychologists plot the points, draw a straight line, and stop. They may draw the line "by eye," or by a formal rule that minimizes the sum of the squared vertical distances of the points about the line. (This book does not use a formal method for exploratory work: It involves too much arithmetic, and inspection of the residuals will allow the analyst to correct errors in an original line.) Tukey points out that this line only describes the general behavior of the data. After a line has been fitted, the interesting work has just begun. To reveal more subtle effects, it is useful to remove the obvious. In this case, one subtracts the y-value of each of the data points from the straight line, and plots these residuals as a function of the x-values. This function will be much flatter than the original function, and it is much easier to see systematic deviations from such a function. This has only occasionally been done in *JEAB* (e.g., Reynolds, 1963). Tukey stresses the value of examining the residuals—subtracting summary values from individual values to look deeper. Successive steps in the analysis lead to improvement in understanding. To make any systematic deviations even more obvious, one can magnify them by expanding the vertical scale.

Myers and Myers (1977) examined the published data of pigeons' performance on concurrent variable-interval schedules. The data from each animal were plotted on the standard coordinates: response proportion on the vertical axis, and reinforcement proportion on the horizontal axis. The data were fit with three functions: $y = x$, $y = ax + b$, and a polynomial. The conclusion was that the slope of the best linear function was less than 1.0 (undermatching), but that the nonlinear function (with more parameters) was best. An alternative approach to these data would be to subtract the matching line and inspect the residuals, then subtract the undermatching line and inspect these residuals. These residuals are not random; there seems to be something special about the extreme cases, i.e., there is (of course) no undermatching at the extremes.

This method works for linear relationships, but what should one do if the relationship between two variables is nonlinear? One possibility is to try to identify the formal rule, and fit the data points to this function. For example, Pierrel (1963) fit a power function relating response rate to change in dB from S+. Tukey suggests that the function be straightened before it is flattened. For example, it is now standard to plot power functions on log-log coordinates. To choose an appropriate transformation, one can simply deal with three points—the two end points and one in the middle. By comparing the slope of the points 1 and 2 with points 2 and 3, one can determine whether the function is changing in an accelerating or decelerating manner, and this suggests approximately which transformation on Tukey's ladder (see below) would straighten the data points. Since only three points need to be checked, it is not time-consuming to try several alternatives. When the data are straightened, the method for plotting the residuals and expanding the scale described in the preceding paragraph can be employed.

Finally, there are complex relationships that cannot be transformed into a straight line. Variability can often obscure a true relationship, and, conversely, it may allow an analyst to mislead the reader into believing a regularity exists when it does not. (A wide line is often used so that the line can be close to the points.) Tukey's methods for smoothing data are much better than unconstrained methods for finding regularities when they exist and only when they exist.

Transformations

Tukey deals extensively with scale transformations. Some psychologists are unwilling to transform their numbers for any reason. They apparently feel that the units that they happened to use to record data are fundamental and should not be changed. Some investigators perform a transformation only if the new scale has a name, e.g., the logarithmic decibel scale for sound intensity and the reciprocal scale for running speed. Others perform a transformation only if it is conventional, e.g., the logarithm of the concentration of a drug. Tukey expands the motivation for transforming data. He provides many examples in which communication is far more exact in a transformed scale than in the original scale. He gives three main reasons for transformations: (a) A transformation may be selected to produce a symmetrical distribution, (b) it may increase the similarity of the spread of different sets of numbers, and (c) it may straighten out a line. He thinks of these as of minor importance, middling importance, and great importance, respectively. Often, a particular transformation will make a single distribution more symmetrical, make the spread of several distributions more similar, and make a line connecting the medians of the distributions straighter. It is rare for a transformation to improve some of these measures and make the others worse.

Some transformations used in articles in *JEAB* are simply operations that make the measures relative, rather than absolute. For example, response rate, the discrimination index, suppression ratio, IRTs per opportunity, and proportion of responses simply express a number as a ratio. They have some interesting and useful properties, but they are not the transformations dealt with in this book. The transformations discussed here range on Tukey's ladder from x^n, x^{n-1}, . . . , x^3, x^2, x, log x, $-1/x$, $-1/x^2$, $-1/x^3$, . . . , $-1/x^n$. Note that the exponents are in the series n, $n-1$, . . . , 3, 2, 1, and -1, -2, -3, . . . , $-n$. The gap between 1 and -1 is not filled by 0—it would not be reasonable to transform all numbers to a single number. The remarkable fact is that log x fills the gap, right between the positive and negative exponents. Movement from left to right on the ladder of transformations emphasizes differences among the smaller numbers; movement from right to left emphasizes differences among the larger numbers.

Transformations of this sort have been used very sparingly in *JEAB*. Response rate is normally reported in arithmetic units, but occasionally it is scaled in logarithmic units (e.g., Nevin, 1974). The logarithm of the control rate is a standard measure in drug studies. The logarithm of the drug rate is another plausible measure, and the relationship of the log control rate and log drug rate would provide a visual representation of the rate dependency hypothesis. The typical "drug effect" measure is far too complicated, as Gonzalez and Byrd (1977) report, since it requires the reader to distinguish between the observed slope and a slope of -1. It is far easier to see small deviations from a slope of 0.

Other dependent variables, e.g., counts and latencies, are occasionally transformed by taking the square root, the logarithm, or the reciprocal. Various independent variables are also sometimes scaled in logarithmic units, e.g., time intervals and drug concentrations. It would be desirable for the analyst to choose the data representation on some rational basis. Unfortunately, articles in *JEAB* that use a transformation seldom make the purpose of the transformation explicit, so one can only guess whether or not the scale used was chosen on a rational basis.

The range of plausible transformations is much greater than the range typically employed in *JEAB* (from square root through logarithm to the reciprocal). In one chapter, Tukey describes unusual transformations (folded logarithms and folded square roots) that often straighten functions based on counted data. This would include psychophysical data in which the percentage response is related to a stimulus variable, usually as an S-shaped function. One virtue of

straightening such data is that it makes it possible to use a measure of discriminability that includes all of the data.

Does the transformation to straighten the data "distort" it? Not really. To say that a function is linear on a log-log scale does not "distort" a power function; it is simply another description of it. It is usually a description to be preferred since it is easier for us to think in terms of straight lines than other functions.

Tables

Tukey develops in great detail important methods for analyzing tables in which there is one response measure for each combination of two or more conditions. For example, Shimp and Moffitt (1977) describe tests of short-term memory of pigeons for tilted lines. The probability of a correct choice was the response measure. In one experiment the conditions were (a) the retention interval of .1, 1.0, 2.0, 4.0, 8.0, 16.0, 24.0, and 32.0 sec, and (b) the events during the retention interval: a houselight on, interfering lines on the center key, both houselight and interfering lines, and neither houselight nor interfering lines. The investigators show that the probability of a correct choice decreases in a fairly linear way with the logarithm of the retention interval, and their inspection of the figures suggests that recall was best without houselight or interfering lines, worst with both, and intermediate with one or the other. These data are in a form for the two-way analysis proposed by Tukey. Although there are many options, the general idea is to examine the medians of the rows and columns, subtract them out, and examine the residuals. The process can be repeated until the residual table contains only (one would hope) small random variation, and the main effects can then be identified. These effects and the residuals can then be displayed graphically. It is useful to see how much more informative this analysis can be than one that deals only with the row and column medians without iteration or examination of residuals.

Shimp and Moffitt (1977) reported data from four pigeons, and a mean. The methods for analyzing tables can be applied when one of the factors is subject or one of the factors is event during the retention interval. In both cases, these methods are more informative than a group mean or median. All articles in *JEAB* report data from individual subjects: It is an editorial policy inscribed on the masthead. In addition, however, investigators might consider Tukey's methods as an alternative to a mean or a median. They make it possible to investigate individual differences in performance if an adequate number of animals is tested. It would seem that a description of individual differences would be appropriate in a journal that "is primarily for the original publication of experiments relevant to the behavior of individual organisms." If Tukey's methods had been available earlier, there might now be less concern over conclusions about groups that are not true for individuals.

The problem of combining subjects is just one case of the more general problem of combining any data. For example, if each of several animals learns a task abruptly but at different times, the mean curve may rise gradually. Similarly, if a single animal on successive fixed-intervals has a period of nonresponse followed by a period of response, but the start of responding begins at different times, the mean curve may rise gradually. In both cases, it is more informative to combine relative to the point of change (backward learning curve and breakpoint analysis, respectively). This is really a problem of defining good variables and, being somewhat specific to different subject matters, this book has little to contribute.

Inferential Statistics

This book serves as a prelude to what is the major subject matter of the typical statistics book: inferential statistics. Tukey's position is that the exploratory phase of data analysis must precede the confirmatory phase. This book contains no inferential statistics, but after the exploratory phase is completed, it is often desirable to attempt to confirm the results.

Very few articles in *JEAB* use any inferential statistics (for example, 10% of the articles published in 1978). There are probably many reasons for this; a few examples are:

1. Some investigators prefer not to go beyond the data. They tell only the facts about any observed sample and avoid all conclusions about the population that was not observed. Such underanalysis of data avoids responsibility. It makes the reader try to do the investi-

gator's work; however, the investigator is in a better position to reach general conclusions.

2. To some investigators, the need to use any statistical test is an admission that there is too much variability. They believe the alternative to statistical analysis is to do experiments with little variability. Of course, it is not necessary to use formal means to test the obvious, but it is often desirable to analyze more than the obvious effects.

3. Inferential statistics can lead to erroneous conclusions. If the assumptions are in error, the conclusions can be wrong. Even if the assumptions are correct, there is the possibility of Type I or Type II errors. One way to avoid the criticism that a statistical test has been misapplied is not to apply it at all: It is no fun to play the game if you don't know the rules. Of course, people who reach conclusions in a more intuitive way can also be wrong, and it is more difficult to estimate the probability of error in these cases.

4. There is a belief, one that is not correct, that a statistical analysis cannot be used on individual subjects.

5. Some individuals may enjoy reaching strong conclusions from weak data. This encourages criticism, and when these folk are challenged, they have the opportunity of demonstrating the correctness of their original conclusions (i.e., of sandbagging).

6. Some psychologists misuse inferential statistics. They may rush into complicated analyses of variance without ever having studied the data; they may get so far from the original data that they report the level of significance but forget to report the direction of the difference; they may fail to use the best experimental design because they do not know how to deal statistically with the results of this design; they may perform the test only because they believe a statistical blessing is desirable.

Tools of Data Analysis

Tukey favors analysis of data with little more than pencil and paper. Specifically, there is no need for a calculator, a computer, or a lettering guide to do the analyses he proposes. The equipment he recommends includes a four-color pen, graph paper with ruling at intervals of 5s and 10s (with the 10s darker), a transparent straight edge, tracing paper, index cards, and a few small tables that are included in the book. It is remarkable how much can be accomplished with such primitive tools, but many psychologists will prefer to use his methods with more advanced tools. His one-page tables for logarithms, square roots, and reciprocals are faster to use than standard tables, but they are not as handy as a calculator with these functions. And a computer is highly desirable, especially for dealing with large data bases and for the iterative procedures that Tukey describes. For some of the more complex procedures (e.g., the iterative analysis of three-way tables), the author also comments on the virtues of a computer. But the emphasis of this book is on hand analysis because the author wants to encourage the data analyst to look at his data and think about it during each step of an analysis, and to proceed in a flexible manner. In the past, many psychologists have used computers only with canned programs. With the reduction in cost and increased availability of minicomputers and microprocessor-based computers, it is now possible to have programs of data analysis under the control of the investigator. For example: For less than $1000, it is now possible to purchase a microprocessor-based computer that can be programmed in BASIC, that permits the investigator immediate feedback from large data files. A graphics scope increases the price, but also the potential, of such a system. The methods described by Tukey can be implemented on such a device or on a time-shared computer, and the use of a computer in this way should greatly increase the productivity of the data analyst.

Generality of Methods

Few of the examples are drawn from psychology, but most of the methods described are applicable to results from psychological experiments. The analyses proposed in this book are data-driven, not theory-driven. Tukey demonstrates that an analyst with no knowledge of the subject matter can, with appropriate methods, discover a great deal in a body of data. If the data analyst is also an expert in the subject matter, however, it is possible that discoveries will be made more quickly and more certainly, since the person will know what to look for. This is particularly likely in a situation involving a very large number of potential factors that might influence the response measure individually,

or only in certain combinations. A general data analyst may search data in an organized fashion for anything; a subject-matter expert is more likely to search for something specific, i.e., to be goal driven.

Influence of the Book

How will methods of data analysis in *JEAB* change? Not everyone will want to learn new methods of data analysis. For some, this approach to data analysis may represent too much work. There are investigators who, after expending an enormous amount of energy in collecting data, spend a trivially small percentage of the total research time looking at the data. Some investigators simply hand the results to a secretary, a research assistant, or to a computer for analysis. When results are given to another for analysis, they are generally accompanied with specific instructions regarding the method of analysis. This is no way to explore data.

Some investigators may be willing to spend the effort to look at the data in a flexible manner, but may think it's cheating—even random data may have some post hoc systematic tendencies. This is not relevant in the discovery phase. For confirmation, it is desirable to replicate. Some investigators may believe there is only one appropriate way to analyze some data, but this belief is due to limited imagination or overreliance on the conventional methods developed by others.

Some investigators will use Tukey's methods to explore data published in *JEAB*. The major question is whether or not they can describe their results with these methods more simply, more completely, or more accurately than with conventional methods. Speculation is not useful. In a few years, there should be examples.

Concluding Remarks

Does the book succeed? It depends on the effort of the reader. The book appears to be easy to read. There are many concrete examples and hardly any mathematical symbols or equations. The arithmetic functions are all simple, and medians are featured. There are, however, difficulties.

1. There is a large, idiosyncratic vocabulary that must be learned. The words are short and friendly, e.g., "batch," "centering," "cutting," "E-trace," "flog," "froot," "hinge," "stem and leaf." They are clearly defined and they add a certain charm to the book, although more standard words would make the book easier to read.

2. Rereading is essential to distinguish between important and incidental ideas. The author tries to help the reader with a didactic style. For example, when he gets to an important point, e.g., "There is no excuse for failing to plot and look," the words appear in bold print (p. 43). It is still difficult to identify the major points on first reading since the amount of space devoted to a topic is not closely related to its importance.

3. The rationale for the methods is not developed here or in the references. In fact, the only two references are to a text by Mosteller and Tukey, and to the bible.

I have read the book, used it in a graduate course, and used some of the methods for analysis of research data. There is a great deal I don't know about it. The best way to understand the material is to follow the examples in the chapters and to do some of the problems at the ends of the chapters. The problems are not difficult, but they are time consuming. They provide practice, and they reveal the power of the methods; it is instructive to discover what one can see after doing a particular analysis that one could not see previously.

In some respects, this emphasis on descriptive statistics is a return to pre-Fisher days when this was all we knew how to do. This book presents some new ones. The modern techniques are important for all empirically minded psychologists to know about. This book is one source, but there are now others (McNeil, 1977; Mosteller & Tukey, 1977). Psychologists who adopt methods of the sort proposed by Tukey and who adopt his general approach toward data analysis may discover more in their data and get greater enjoyment while analyzing their data.

REFERENCES

Blough, D. S. A method for obtaining psychophysical thresholds from the pigeon. *Journal of the Experimental Analysis of Behavior,* 1958, **1,** 31-43.

Clark, F. C. The effect of deprivation and frequency of reinforcement on variable-interval responding. *Journal of the Experimental Analysis of Behavior,* 1958, **1,** 221-228.

Gonzalez, F. A., & Byrd, L. D. Mathematics underlying the rate-dependency hypothesis. *Science,* 1977, **195,** 545-550.

Mechner, F. Probability relations within response sequences under ratio reinforcement. *Journal of the Experimental Analysis of Behavior,* 1958, **1,** 109-121.

McNeil, D. R. *Interactive data analysis.* New York: Wiley, 1977.

Mosteller, F., & Tukey, J. *Data analysis and regression: A second course in statistics.* Reading, Mass.: Addison-Wesley, 1977.

Myers, D. L., & Myers, L. E. Undermatching: A reappraisal of performance on concurrent variable-interval schedules of reinforcement. *Journal of the Experimental Analysis of Behavior,* 1977, **25,** 203-214.

Nevin, J. A. On the form of the relation between response rates in a multiple schedule. *Journal of the Experimental Analysis of Behavior,* 1974, **21,** 237-248.

Pierrel, R. A generalization gradient for auditory intensity in the rat. *Journal of the Experimental Analysis of Behavior,* 1958, **1,** 303-313.

Reynolds, G. S. Some limitations on behavioral contrast and induction during successive discrimination. *Journal of the Experimental Analysis of Behavior,* 1963, **6,** 131-139.

Shimp, C. P., & Moffit, M. Short-term memory in the pigeon: Delayed-pair-comparison procedures and some results. *Journal of the Experimental Analysis of Behavior,* 1977, **28,** 13-25.

Zeiler, M. D. Editorial. *Journal of the Experimental Analysis of Behavior,* 1977, **27,** 1-2.

Received November 30, 1978
Final acceptance December 20, 1978